Smith et al. take up the challenge to assess the performance of combined SMB-FD models against laser-derived observations of elevation change over parts of the Greenland ice sheet where they assume ice-dynamical effects to be negligible. This is important, as it allows us to understand how to improve altimetry-based estimated of GrIS mass balance using firn and SMB models.

The paper is clearly written (in most places), and the scientific analyses are sound (in most places).

My only major critique of this paper is that the analysis of the dh correction, as well as the scaling experiments in section 3.2, are only presented in terms of histograms over the entire ice sheet, or over aggregated sections of the ice sheet.

Much more insight would be provided to the reader if time series were presented from a selection of locations across the ice sheet. What do time series of dh, dhm, dhc, dhFAC and dhSMB look like for an individual location in, for example, the lower western accumulation area, the southern interior, the northern interior, the northeast and the southeast? Rather than having to guess the physical reasons for improved agreement (reduced residuals), it would become clear at a process level from the time series.

As this is my only major concern, I encourage the authors to expand the paper to accommodate for it. It would strengthen further the discussion about the scaling experiments, because the authors will have figures with time series of dhFAC and dhSMB that immediately make obvious why scaling of dhFAC works and that of dhSMB won't.

Line by line comments:

L 29: heights vary -> elevation varies

L 38: perhaps good to clarify that you are referring to a climatologically mean surface mass balance here

L 51: This is a confusing statement. FD models are forced by meteorological parameters as well as by mass fluxes, both of which are computed by an SMB model.

L 68: Here you focus mostly on the densification part of an FD. However, the thermodynamical part of FD models is usually also evaluated against observations of deep temperature.

L 71: In Munneke et al. (2015), laser-observed dh/dt was tested against an FD model at selected locations in order to evaluate their model.

L 93: I assume that the separation of 3.3 km refers to the ground projection of the lasers, not of the lasers themselves

L 105: please reformulate: a strategy does not measure anything.

L 105: “At each of a set of reference points…” this sentence does not flow well

L 129: Why is it safe to assume that the errors derived from release-003 products are not too optimistic?

Table 1: Listing the internal model variables feels redundant since they are never referred to in the manuscript. This table can either be moved to the supplementary materials or removed entirely.

L 220: In 2008, Helsen et al. showed that systematic surface elevation change can be the delayed result of multi-decadal or even centennial variability in SMB. In the present setup of your study, this effect is not accounted for. Rather, like in other studies, changes are defined with respect to a reference period (in your case, 1980-1995) over which no change is assumed. However, in the interior over which you evaluate the SMB/FD models, any residual between observations and models could be caused by these very long-term effects originating from quite deep in the firn.

L 305 (figure 2): see major issue above. The 32 tiled maps are a very comprehensive way of presenting the data, but it lacks in detail, making it hard to judge the models against observations at key locations. My suggestion would be to add a figure with time series of dh, dhm and dhc for a few selected locations (e.g. west coast, southern interior, northern interior, NE coast, SE coast). In that way, it becomes much easier to appreciate the temporal simulation of elevation change by the models compared to the observations.

L 330 (figure 4): perhaps clarify here that the scaling factors X were defined such that dh - X dhm == 0

L 345: what does the scaling imply? Is surface density not sufficiently captured? Is there a structural overestimation of snowfall and/or melt? Is there a structural error in the ICESat observations?

L 425 (figure 7): the effect of only scaling dhFAC (light green line) is invisible in the graph.

L 425 (figure): why does rescaling the dhSMB make almost no difference, as opposed to rescaling dhFAC? Please elaborate on this.

L 456: why does it help to isolate errors in high-elevation melt when the agreement at lower elevations is good? Can we simply assume that an SMB model will perform well at higher elevations (snow albedo dominated) when it does so at lower elevations (ice albedo dominated)?

L 471: The elevation change in GFSCv1.2 is much less sensitive to melt events than the other two models. At the same time, its surface density has increased to 327-387 kg/m3, which is higher than the mean 315 kg/m3 reported by Fausto et al. in 2018. The surface elevation change associated with a melt event is approximately the amount of melt per unit area divided by the density of the melted snow: $\Delta h_{melt} \approx \Delta m_{melt}/\rho_{surface}$. Why do you think the higher surface density cannot explain the lower sensitivity of GFSCv1.2 to melt events?

L 514: GFSCv1.1 and GFSCv1.1 -> GFSCv1.1 and GFSCv1.2